AD-A198 321

DTIC FILE COPY

2

PAGE (When Data Entered)

MENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFOSR-TR- 88-0788	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Weighted and clouded distributions	5. TYPE OF REPORT & PERIOD COVERED ~~Technical~~ Reprint - February 1988
	6. PERFORMING ORG. REPORT NUMBER 88-01
7. AUTHOR(s) C. Radhakrishna Rao	8. CONTRACT OR GRANT NUMBER(s) AFOSR Grant ~~AFSO~~-88-0030
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Multivariate Analysis Fifth floor - Thackeray Hall University of Pittsburgh, Pittsburgh, PA 15260	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304 A6
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Department of the Air Force Bolling Air Force Base, DC 20332	12. REPORT DATE February 1988 / 13. NUMBER OF PAGES 38
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) same as 11	15. SECURITY CLASS. (of this report) Unclassified / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC ELECTE AUG 25 1988 S D E

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

damage model, probability sampling, quadrat sampling, size biased sampling, truncation, weighted distribution

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

The concept of weighted distributions can be traced to the study of *effects of methods of ascertainment upon the estimation of frequencies* by Fisher in 1934. It was formulated in general terms by the author in a

Continued--

DD FORM 1 JAN 73 1473

88 8 25 139

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD-A198 321

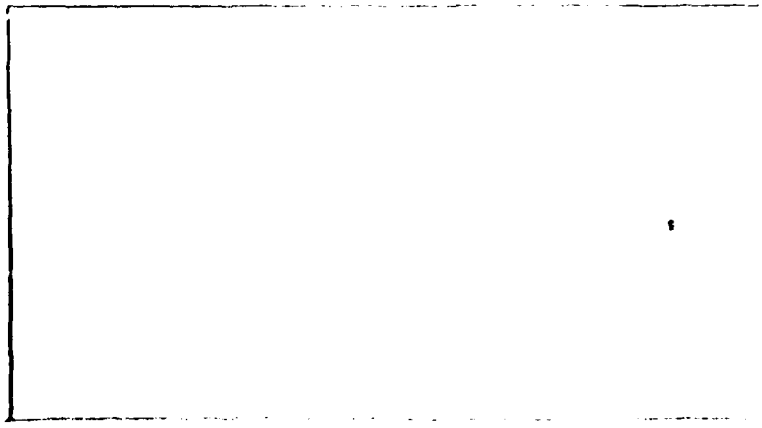
20 Abstract (continued)

paper presented at the First International Symposium on Classical and Contagious Distributions held in Montreal in 1963. Since then a number of papers have appeared on the subject. This article reviews the previous work and the current developments with some examples.

Weighted distributions occur in a natural way when adjustments have to be made in the original probability distribution due to deviations from simple random sampling in collecting data, as when the events that occur do not have the same chance of coming into the sample. The examples include: p.p.s. (probability proportional to size) sampling in sample surveys, damage models, visibility bias in quadrat sampling in ecological studies, sampling through effected individuals in genetic studies, waiting time paradox and so on.

88-0030

AFOSR-TR- 88-0788



Center for Multivariate Analysis

University of Pittsburgh

WEIGHTED AND CLOUDED DISTRIBUTIONS*

C. Radhakrishna Rao

University of Pittsburgh
Pittsburgh, PA 15260

Technical Report No. 88-01

February 1988

Center for Multivariate Analysis
5th Floor Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

* Research sponsored by the Air Force Office of Scientific Research under Grant AFSO-88-0030. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

WEIGHTED AND CLOUDED DISTRIBUTIONS*

C. Radhakrishna Rao

University of Pittsburgh
Pittsburgh, PA 15260

ABSTRACT

The concept of weighted distributions can be traced to the study of *effects of methods of ascertainment upon the estimation of frequencies* by Fisher in 1934. It was formulated in general terms by the author in a paper presented at the First International Symposium on Classical and Contagious Distributions held in Montreal in 1963. Since then a number of papers have appeared on the subject. This article reviews the previous work and the current developments with some examples.

Weighted distributions occur in a natural way when adjustments have to be made in the original probability distribution due to deviations from simple random sampling in collecting data, as when the events that occur do not have the same chance of coming into the sample. The examples include: p.p.s. (probability proportional to size) sampling in sample surveys, damage models, visibility bias in quadrat sampling in ecological studies, sampling through effected individuals in genetic studies, waiting time paradox and so on.

AMS 1980 Subject Classifications: Primary 60E05, secondary 62D05.

Key words and phrases: damage model, probability sampling, quadrat sampling, size biased sampling, truncation, weighted distribution.

* Research sponsored by the Air Force Office of Scientific Research under Grant AFSO-88-0030. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

DEDICATION

Professor K. Nagabhushanam was one of the two inspiring teachers I had when I was pursuing my graduate studies in mathematics at the Andhra University, Waltair. He and Professor V. Ramaswamy not only taught us mathematics but also prepared us to think in terms of mathematics and to use mathematics as an abstract logical method in solving complex problems in any field of inquiry. This training was a great asset to me when I started on my research career. I had kept in touch with Professor Nagabhushanam after I left Andhra University as he was keenly interested in my activities and often encouraged me in my research work. It is, indeed, a great honor to contribute to the memorial volume of my respected teacher. The contents of this chapter are specifically addressed to the students and teachers of statistics who are looking for simple examples to demonstrate some natural pitfalls in statistical data analysis and inference.

1. INTRODUCTION

In statistical inference, i.e., making statements about a population on the basis of a sample drawn from it, it is necessary to identify the holy trinity, viz., the sample space Ω, Borel field of sets F defined on Ω and a family of probability measures P defined on F. Statistical analysis is concerned with setting up a correspondence between a sample (a member of Ω) and an element (or subset of elements) of P. An important part of the trinity is the specification P. Wrong specification may lead to wrong inference, which is sometimes called the *third kind of error* in statistical parlance.

The problem of specification is not a simple one. A detailed knowledge of the procedure actually employed in acquiring data is an essential ingredient in arriving at a proper specification. The situation is more complicated with field observations and nonexperimental data, where nature produces events according to a certain stochastic model, and the events are observed and recorded by field investigators. There does not always exist a suitable sampling frame for designing a sample survey to ensure that the events which occur have specified (usually equal) chances of coming into the sample. In practice, all the events that occur in nature cannot be brought into the sample frame. For instance, certain events may not be observable and therefore missed in the record. This gives rise to what are called truncated, censored or incomplete samples. Or an event that has occurred may be observable only with a certain probability depending on the nature of the event, such as its conspicuousness and the procedure employed to observe it, resulting in unequal probability sampling. Or an event which has occurred may change in a random way by the time or during the process of observation so that what comes on record is a modified event, in which

case the change or damage has to be appropriately modeled for statistical analysis. Sometimes, events from two or more different sources having different stochastic mechanisms may get mixed up and brought into the same record, resulting in contaminated samples. In all of these cases, the specification for the original events (as they occur) may not be appropriate for the events as they are recorded (observed data) unless it is suitably modified. Examples of such situations are given in Rao (1965, 1975, 1985).

In a classical paper, Fisher (1934) demonstrated the need for such an adjustment in specification depending on the way data are ascertained. The author extended the basic ideas of Fisher in Rao (1965) and developed the theory of what are called weighted distributions as a method of adjustment applicable to many situations. In this paper we discuss the general theory and some recent developments through some examples.

2. TRUNCATION

Some events, although they occur, may be unascertainable, so that the observed distribution is truncated to a certain region of the sample space. For instance, if we are investigating the distribution of the number of eggs laid by an insect, the frequency of *zero eggs* is not ascertainable. Another example is the frequency of families where both parents are heterozygous for albinism but have no albino children. There is no evidence that the parents are heterozygous unless they have an albino child, and the families with such parents and having no albino children get confounded with normal families having no children. The actual frequency of the event *zero albino children* is thus not ascertainable.

In general, if $p(x,\theta)$ is the p.d.f. (probability density function for a continuous variable or probability for a discrete variable), where θ denotes

an unknown parameter, and the random variable X is truncated to a specified region $T \subset \Omega$ of the sample space, then the p.d.f. of the truncated random variable X^T is

$$p^T(x,\theta) = \frac{w(x,T)p(x,\theta)}{u(T,\theta)} \tag{2.1}$$

where $w(x,T) = 1$ if $x \in T$ and $= 0$ if $x \notin T$, and $u(T,\theta) = E[w(X,T)]$. The expression (2.1) is the original probability density weighted by a suitable function, and it provides a simple example of a weighted probability distribution whose general definition is given in the next section.

Suppose the event zero is not observable in sampling from a binomial distribution with index n and probability of success π. Let R^T denote the TB (truncated binomial) random variable. Then

$$P(R^T = r) = \frac{\binom{n}{r}\pi^r(1-\pi)^{n-r}}{1 - (1-\pi)^n}, \quad r = 1,\ldots,n. \tag{2.2}$$

For such a distribution

$$E(R) = \frac{n\pi}{1 - (1-\pi)^n} \quad \text{and} \quad E\left(\frac{R}{n}\right) = \frac{\pi}{1 - (1-\pi)^n} \tag{2.3}$$

which are somewhat larger than those for a complete binomial, for which the above values are $n\pi$ and π respectively.

The following data relate to the numbers of brothers and sisters in families of the girls whose names were found in a private telephone notebook of a European professor. (The first number within the brackets gives the number of sisters including the respondent and the second number, that of her brothers.

$$\begin{gathered}(1,0),(1,0),(1,1),(1,1),(1,1),(1,1),(1,1),(1,1),(1,1),(1,1)\\(1,1),(2,0),(2,0),(2,0),(2,1),(2,1),(2,1),(2,1),(1,2),(1,2)\\(3,0),(3,1),(3,1),(1,3),(1,3),(4,0),(4,1),(1,4)\end{gathered} \tag{2.4}$$

Since at least one girl is present in the family, we may try and see whether the data conform to a TB distribution with the observation on *zero sisters* missing. The expected number of girls under this hypothesis, assuming $\pi = 0.5$, is

$$\sum_{n=1}^{5} f(n)E(r|n) \tag{2.5}$$

where f(n) is the observed number of families with size n (i.e., the total number of brothers and sisters). Using the formulas (2.3) and (2.5) and the data (2.4), we have:

Number of	observed	expected
Sisters	47	46
Brothers	30	31

The observed figures seem to be in good agreement with those expected under the hypothesis of truncated binomial. However, a different story may emerge in a similar situation as in the following data giving the numbers of sisters and brothers in the families of girl acquaintances of a male student in Calcutta.

$$(2,1), (1,1), (3,0), (2,0), (3,1), (1,0), (2,1), (1,0), (1,1), (1,1). \tag{2.6}$$

The expected numbers of sisters under the hypothesis of truncated binomial is 9.5 (using the formulas (2.3) and (2.5)) whereas the observed number is 17. The truncated binomial is not appropriate for the data (2.6) and it appears that the mechanisms of encountering girls seem to be different in the cases of the professor and the student.

Note that if we sample a number of households in a city and ascertain the numbers of brothers and sisters (i.e., sons and daughters) in each household, then we expect the number of sisters to follow a complete binomial distribution. If from such data we omit the households which do not

have girls, then the data would follow a truncated binomial distribution. We shall see in the next section that a different distribution holds when data are ascertained about the sibs from a sample of boys or girls one *encounters*. The case of the student seems to fall in such a category.

3. WEIGHTED DISTRIBUTIONS

In Section 2, we have considered a situation where certain events are unobservable. But a more general case is when an event that occurs has a certain probability of being recorded (or included in the sample). Let X be a random variable with $p(x,\theta)$ as the p.d.f., where θ is a parameter, and suppose that when $X = x$ occurs, the probability of recording it is $w(x,\alpha)$ depending on the observed x and possibly also on an unknown parameter α. Then the p.d.f. of the resulting random variable X^w is

$$p^w(x,\theta,\alpha) = \frac{w(x,\alpha)p(x,\theta)}{E[w(x,\alpha)]}. \tag{3.1}$$

Although in deriving (3.1) we chose $w(x,\alpha)$ such that $0 \leq w(x,\alpha) \leq 1$, we may formally define (3.1) for any arbitrary nonnegative function $w(x,\alpha)$ for which $E[w(x,\alpha)]$ exists. The p.d.f. so obtained is called a weighted version of $p(x)$ and denoted by $p^w(x)$. In particular the weighted distribution

$$p^w(x,\theta) = \frac{f(x)p(x,\theta)}{E\big(f(x)\big)} \tag{3.2}$$

where $f(x)$ is some monotonic function of x, is called a size biased distribution. When x is univariate and nonnegative, the weighted distribution

$$p^w(x,\theta) = \frac{x^\alpha p(x,\theta)}{E(x^\alpha)} \tag{3.3}$$

introduced in Rao (1965) has found applications in many practical problems

(see Rao (1985)). When $\alpha = 1$, it is called a length (size) biased distribution. For example, if X has the logarithmic series distribution

$$\frac{\alpha^r}{-r \log(1-\alpha)}, \quad r = 1,2,... \tag{3.4}$$

then the distribution of the length biased variable is

$$(1-\alpha)\alpha^{n-1}, \quad r = 1,2,...$$

which shows that $X^w - 1$ has a geometric distribution. A truncated geometric distribution is sometimes found to provide a good fit to an observed distribution of family size (Feller, 1968). But, if the information on family size has been ascertained from school children, then the observations may have a size biased distribution. In such a case, a good fit of the geometric distribution to the observed family size would indicate that the underlying distribution is, in fact, a logarithmic series.

Table 1 gives a list of some basic distributions and their size biased forms. It is seen that the size biased form belongs to the same family as the original distribution in all cases except the logarithmic series.

An extensive literature on weighted distributions has appeared since the concept was formalized in Rao (1965); it is reviewed with a large number of references in a paper by Patil (1984) with special reference to the earlier contributions by Patil and Rao (1977, 1978) and Patil and Ord (1976). Rao (1985) contains an updated review of the previous work and some new results.

Table 1. Certain Basic Distributions and Their Size-Biased Forms

Random variable (rv)	pf (pdf)	Size-biased rv
Binomial, $B(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$1 + B(n-1,p)$
Negative binomial, $\mathrm{NB}(k,p)$	$\binom{k+x-1}{x}q^x p^k$	$1 + \mathrm{NB}(k+1,p)$
Poisson, $\mathrm{Po}(\lambda)$	$e^{-\lambda}\lambda^x/x!$	$1 + \mathrm{Po}(\lambda)$
Logarithmic series, $L(\alpha)$	$\{-\log(1-\alpha)\}^{-1}\alpha^x/x$	$1 + \mathrm{NB}(1,\alpha)$
Hypergeometric, $H(n,M,N)$	$\binom{n}{x}\frac{M^x(N-M)^{n-x}}{N^n}$	$1 + H(n-1,M-1,N-1)$
Binomial beta, $\mathrm{BB}(n,\alpha,\gamma)$	$\binom{n}{x}\frac{\beta(\alpha+x,\gamma+n-x)}{\beta(\alpha,\gamma)}$	$1 + \mathrm{BB}(n-1,\alpha,\gamma)$
Negative binomial beta, $\mathrm{NBB}(k,\alpha,\gamma)$	$\binom{k+x-1}{x}\frac{\beta(\alpha+x,\gamma+k)}{\beta(\alpha,\gamma)}$	$1 + \mathrm{NBB}(k+1,\alpha,\gamma)$
Gamma, $G(\alpha,k)$	$\alpha^k x^{k-1}e^{-\alpha x}/\Gamma(k)$	$G(\alpha,k+1)$
Beta first kind, $B_1(\delta,\gamma)$	$x^{\delta-1}(1-x)^{\gamma-1}/\beta(\delta,\gamma)$	$B_1(\delta+1,\gamma)$
Beta second kind, $B_2(\delta,\gamma)$	$x^{\delta-1}(1+x)^{-\gamma}/\beta(\delta,\gamma-\delta)$	$B_2(\delta+1,\gamma-\delta-1)$
Pearson type V, $\mathrm{Pe}(k)$	$x^{-k-1}\exp(-x^{-1})/\Gamma(k)$	$\mathrm{Pe}(k-1)$
Pareto, $\mathrm{Pa}(\alpha,\gamma)$	$\gamma\alpha^\gamma x^{-(\gamma+1)}$, $x \geq \alpha$	$\mathrm{Pa}(\alpha,-1)$
Lognormal, $\mathrm{LN}(\mu,\sigma^2)$	$(2\pi\sigma^2)^{-1/2}x^{-1}\exp-\left(\frac{\log x-\mu}{\sigma\sqrt{2}}\right)^2$	$\mathrm{LN}(\mu+\sigma^2,\sigma^2)$

4. p.p.s. SAMPLING

An example of a weighted distribution arises in sample surveys when unequal probability sampling or what is known as p.p.s. (probability proportional to size) sampling is employed. A general version of the sampling scheme involves two random varianbes X and Y with p.d.f. $p(x,y,\theta)$ and a weight function $w(y)$ which is a function of y only, giving a weighted p.d.f.

$$p^w(x,y,\theta) = \frac{w(y)p(x,y,\theta)}{E[w(Y)]} . \tag{4.1}$$

In sample surveys, we obtain observations on (X^w,Y^w) from the p.d.f. (4.1) and draw inference on the parameter θ.

It is of interest to note that the marginal p.d.f. of X^w is

$$p^w(x,\theta) = \frac{w(x,\theta)p(x,\theta)}{E[w(X,\theta)]} \tag{4.2}$$

which is a weighted version of $p(x,\theta)$ with the weight function

$$w(x,\theta) = \int p(y|x)w(y)dy. \tag{4.3}$$

If we have a sample of size n

$$(x_1,y_1), \ldots, (x_n,y_n) \tag{4.4}$$

from the distribution (4.1), then an estimate of E(X), the mean with respect to the original p.d.f. $p(x,y,\theta)$, which is the parameter of interest, is

$$\frac{E[w(Y)]}{n} \sum_{i=1}^{n} \frac{x_i}{w(y_i)} \tag{4.5}$$

which is an unbiased estimator of E(X). The estimator

$$\frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.6}$$

would be an unbiased estimator of $E(X^W)$, the mean with respect to weighted p.d.f. $p^W(x,\theta)$ as in (4.3).

5. WEIGHTED BINOMIAL DISTRIBUTION: TWO EMPIRICAL THEOREMS

Suppose that we ascertain from each male member in a class or in any congregation the number of brothers including himself and the number of sisters he has and raise the following question. What is the approximate value of B/(B+S), where B and S are the total numbers of brothers and sisters in all the families of the male members? It is clear that we are sampling from a truncated distribution of families with at least one male member so that B/(B+S) should be larger than one half. But by how much? Surprisingly, when k, the number of males asked, is not very small, one can make accurate predictions of the relative magnitudes of B and S, and of the ratio B/(B+S). This may be stated in the form of an empirical theorem.

Empirical Theorem 1: Let k male members observed in any gathering have a total number of B brothers (including themselves) and a total number of S sisters. Then the following predictions can be made:

(i) B is much greater than S.

(ii) B - k is approximately equal to S.

(iii) B/(B+S) is larger than one half. It will be closer to

$$\frac{1}{2} + \frac{k}{2(B+S)} .$$

(iv) (B-k)/(B+S-k) is close to half.

The roles of B and S are reversed if the data are ascertained from the female members in a gathering.

Consider a family with n children. Then on the assumption of a binomial distribution with $\pi = 1/2$ and index n, the probability of r male children is

$$p(r) = \binom{n}{r}\left(\frac{1}{2}\right)^n, \quad r = 0,1,2,\ldots . \tag{5.1}$$

In our case, there is at least one male child so that the appropriate distribution is a truncated one. One possibility is a truncated binomial (TB),

$$p^T(r) = \frac{\binom{n}{r}\left(\frac{1}{2}\right)^n}{1-\left(\frac{1}{2}\right)^n}, \quad r = 1,2,\ldots \tag{5.2}$$

and another is a size biased binomial (WB)

$$p^W(r) = \frac{r\binom{n}{r}\left(\frac{1}{2}\right)^n}{(n/2)} = \binom{n-1}{r-1}\left(\frac{1}{2}\right)^{n-1}, \quad n = 1,2,\ldots . \tag{5.3}$$

In Rao (1977), it was argued that (5.3) is more appropriate for the observed data than (5.2). Table 2 gives the observed frequency distributions of the number of brothers in families of different sizes based on the data collected separately from the male and female students in the universities at Shanghai (China), Manila (Phillipines), and Bombay (India), and the expected values on the hypotheses of TB as in (5.2) and WB as in (5.3).

Table 2. Observed frequencies of the number of brothers in families of different sizes and expected frequencies under the hypotheses of TB and WB distributions.

(Data from male students in Shanghai, Manila and Bombay)

	n = 1			n = 2			n = 3		
No. of brothers	observed	expected TB	expected WB	observed	expected TB	expected WB	observed	expected TB	expected WB
1	6	6	6	24	28.7	21.5	12	20.1	11.7
2				19	14.3	21.5	24	20.1	23.5
3							11	6.7	11.7
TOTAL	6	6	6	43	43.0	43.0	47	46.9	46.9

	n = 4			n = 5			n = 6		
No. of brothers	observed	expected TB	expected WB	observed	expected TB	expected WB	observed	expected TB	expected WB
1	8	11.2	5.3	5	6.5	2.5	1	1.9	0.6
2	10	16.8	15.7	8	12.9	10.0	4	4.8	3.1
3	17	11.2	15.7	15	12.9	15.0	4	6.3	6.3
4	7	2.8	5.3	10	6.5	10.0	9	4.8	6.3
5				2	1.3	2.5	2	1.9	3.1
6							0	0.3	0.6
TOTAL	42	42.0	42.0	40	40.1	40.0	20	20.0	20.0

It is seen from the above table that the WB (weighted binomial) provides a better fit than the TB (truncated binomial) indicating that a family with r brothers is sampled with probability proportional to r.

Accepting the hypothesis of the weighted (size biased) binomial, viz.,

$$p(r) = \binom{n-1}{r-1}\left(\frac{1}{2}\right)^{n-1}, \quad n = 1,2,\ldots,n, \tag{5.4}$$

we immediately find that

$$E(r|n) = \sum_{r=1}^{n} r\binom{n-1}{r-1}\left(\frac{1}{2}\right)^{n-1} = \frac{n+1}{2} \Rightarrow E(r-1) = \frac{n-1}{2}. \tag{5.5}$$

If (r_1,n_1), ..., (r_k,n_k) are observed data with $B = r_1 + \ldots + r_k$, $T = n_1 + \ldots + n_k$ and $S = T - B$, then for given T

$$E(B-k) = \sum_{1}^{k} E(n_i-1) = \sum_{1}^{k} \frac{n_i-1}{2} = \frac{T-k}{2} = E(S). \tag{5.6}$$

$$E(B) = \frac{T+k}{2}, \quad E\left(\frac{B}{T}\right) = E\left(\frac{B}{B+S}\right) = \frac{1}{2} + \frac{k}{2(B+S)}. \tag{5.7}$$

Removing the expectation signs in (5.6) and (5.7), we can assert approximate equalities as stated in Empirical Theorem 1.

During the last twenty years, while lecturing to students and teachers in different parts of the world, I collected data on numbers of brothers and sisters in the family of each individual in my audience. The results are summarized in Tables 3, 4 and 5. It is seen that the predictions as given in the empirical theorem are true in practically every case. As a further test of the weighted binomial, the statistic

$$x^2 = \frac{4([B-k] - [(T-k)/2])^2}{(T-k)} \tag{5.8}$$

which is asymptotically distributed as Chi-square on one degree of freedom is calculated in each case. The Chi-squares are all small providing evidence in favor of the weighted binomial distribution. [Actually, the Chi-squares are too small which needs further study of the mechanism generating the observed data.]

The situation is slightly different in Table 5 relating to the data on

professors. The estimated proportion is more than half in each case, and the Chi-square values are high; this implies that the weight function appropriate to these data is of a higher order than n, the number of brothers. Male professors seem to come from families where sons are disproportionately more than the daughters!

Table 3. Data on Male Respondents (Students)[a]

Place and year	k	B	S	$\frac{B}{B+S}$	$\frac{B-k}{B+S-k}$[b]	χ^2
Bangalore (India, 75)	55	180	127	.586	.496	.02
Delhi (India, 75)	29	92	66	.582	.490	.07
Calcutta (India, 63)	104	414	312	.570	.498	.04
Waltair (India, 69)	39	123	88	.583	.491	.09
Ahmedabad (India, 75)	29	84	49	.632	.523	.35
Tirupati (India, 75)	592	1902	1274	.599	.484	.50
Poona (India, 75)	47	125	65	.658	.545	1.18
Hyderabad (India, 74)	25	72	53	.576	.470	.36
Tehran (Iran, 75)	21	65	40	.619	.500	.19
Isphahan (Iran, 75)	11	45	32	.584	.515	.06
Tokyo (Japan, 75)	50	90	34	.725	.540	.49
Lima (Peru, 82)	38	132	87	.603	.519	.27
Shanghai (China, 82)	74	193	132	.594	.474	.67
Columbus (USA, 75)	29	65	52	.556	.409	2.91
College St. (USA, 76)	63	152	90	.628	.497	.01
Total	1206	3734	2501	.600	.503	0.14

[a] k = number of students, B = total number of brothers including the respondent, S = total number of sisters.

[b] Estimate of π under size biased binomial distribution.

Table 4. Data on Female Respondents (Students)

Place and year	k	B	S	$\frac{S}{B+S}$	$\frac{S-k}{B+S-k}$	χ^2
Lima (Peru, 82)	16	37	48	.565	.464	.36
Los Banos (Philippines, 83)	44	101	139	.579	.485	.18
Manila (Philippines, 83)	84	197	281	.588	.500	.00
Bilbao (Spain, 83)	14	19	35	.576	.525	.10
Shanghai (China, 82)	27	28	55	.662	.500	.00

Table 5. Data on Male Respondents (Professors)

Place and year	k	B	S	$\frac{B}{B+S}$	$\frac{B-k}{B+S-k}$	χ^2
State College (USA, 75)	28	80	37	.690	.584	2.53
Warsaw (Poland, 75)	18	41	21	.660	.525	2.52
Poznan (Poland, 75)	24	50	17	.746	.567	1.88
Pittsburgh (USA, 81)	69	169	77	.687	.565	2.99
Tirupat[illegible]ndia, 76)	50	172	132	.566	.480	0.39
Maracaibo (Venezuela, 82)	24	95	56	.629	.559	1.77
Richmond (USA, 81)	26	57	29	.663	.517	0.03
Total	239	664	369	.642	.535	3.95

Note 1. From (5.7), the expected value of the ratio B/(B+S) for given average family size f = (B+S)/k is as follows for different values of f:

f:	1	2	3	4	5	6
$E(\frac{B}{B+S})$:	1	.75	.67	.625	.6	.58

These figures show that in any given situation where the average family size is not likely to exceed 6, the following predictions can be made about the total number of brothers (B) and of sisters (S) ascertained from the male members in any gathering:

(i) B is much greater than S.

(ii) B/(B+S) is closer to 0.6 or even 2/3 rather than to 1/2.

Surprisingly, these predictions hold even if k, the number of males in a gathering, is small. [This will be a good classroom exercise or a demonstration piece in any gathering. One can make these predictions in advance and demonstrate the accuracy of predictions after collecting the data from male (or female) members.]

Note 2. The probabilities for B > S, B = S, B < S in the case of a weighted binomial distribution for n = 1,2,... are given in Table 6.

Table 6. Probabilities of $B > S$, $B = S$ and $B < S$

n	1	2	3	4	5	6	7	8	9	10
$B > S$	1	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{11}{16}$	$\frac{1}{2}$	$\frac{42}{64}$	$\frac{1}{2}$	$\frac{163}{256}$	$\frac{1}{2}$
$B = S$	0	$\frac{1}{2}$	0	$\frac{3}{8}$	0	$\frac{10}{32}$	0	$\frac{35}{128}$	0	$\frac{126}{256}$
$B < S$	0	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{6}{32}$	$\frac{22}{64}$	$\frac{29}{128}$	$\frac{93}{256}$	$\frac{121}{512}$

It is seen that $P(B > S)$ is much larger than $P(B < S)$ for each n so that in any given audience, the ratio of b_g (males belonging to families with $B > S$) to b_ℓ (those with $B > S$) is likely to be large, depending on the distribution of family sizes. We may now state another empirical theorem.

Empirical Theorem 2. The numbers b_g and b_ℓ are approximately in the ratio of

$$E(b_g) = p_1 + \frac{3}{4} p_3 + \frac{11}{16} p_5 + \ldots + \frac{1}{2}(p_2 + p_4 + \ldots), \tag{5.9}$$

to

$$E(b_\ell) = \frac{1}{4} p_3 + \frac{1}{8} p_4 + \ldots, \tag{5.10}$$

where p_n is the number of families with n children. In western audiences where the expected family size is small, the ratio $b_g : b_\ell$ is likely to be even larger than 4 : 1 and in oriental audiences larger than 2 : 1, which are quite high compared to 1 : 1. [This phenomenon can be predicted and verified by asking the members of an audience to indicate by show of hands how many belong to the category $B > S$ and how many to $B < S$. This will be a good classroom exercise or a demonstration piece in any gathering.]

Note 3. Let $p(b,n)$ be the probability that a family is of size $N = n$ and the number of brothers $B = b$, and suppose that the probability of selecting such a family is proportional to b. Then

$$p^w(b,n) = \frac{bp(b,n)}{E(B)} = \frac{bp(n)p(b|n)}{E(B)}, \tag{5.11}$$

$$p^w(n) = \frac{E(B|n)}{E(B)}p(n). \tag{5.12}$$

When $p(b|n)$ is binomial

$$\begin{aligned} p^w(n) &= \frac{np(n)}{E(N)} \\ E_w\left(\frac{1}{n}\right) &= \frac{1}{E(N)} \end{aligned} \tag{5.13}$$

so that the harmonic mean of observations $n_1, \ldots, n_k$ on N^w, i.e., from the distribution (5.11) or (5.12),

$$\frac{k}{\sum \frac{1}{n_i}} \tag{5.14}$$

is an estimate of $E(N)$ in the original population. If the form of $p(n)$ is known, then one could write down the likelihood of the sample $n_1, \ldots, n_k$ using the probability function (5.12) and determine the unknown parameters by the method of maximum likelihood.

6. ALCOHOLISM, FAMILY SIZE, AND BIRTH ORDER

Smart (1963, 1964) and Sprott (1964) examined a number of hypotheses on the incidence of alcoholism in Canadian families using the data on family size and birth order of 242 alcoholics admitted to three alcoholism clinics in Ontario. The method of sampling is thus of the type discussed in Section 5.

One of the hypotheses tested was that *larger families contain larger numbers of alcoholics than expected.* The null hypothesis that the number of alcoholics is as expected was interpreted to imply that the observations on family size as ascertained arise from the weighted distribution

$$np(n)/E(n), \quad n = 1,2,..., \tag{6.1}$$

where $p(n)$, $n = 1,2,...$, is the distribution of family size in the general population. Smart and Sprott used the distribution of family size as reported in the 1931 census of Ontario for $p(n)$ in their analysis. It is then a simple matter to test whether the observed distribution of family size in their study is in accordance with the expected distribution (6.1).

It may be noted that the distribution (6.1) would be appropriate if we had chosen individuals (alcoholic or not) at random from the general population (of individuals) and ascertained the sizes of the families to which they belonged. But it is not clear whether the same distribution (6.1) holds if the inquiry is restricted to alcoholic individuals admitted to a clinic, as assumed by Smart and Sprott. This could happen, as demonstrated below, under an interpretation of their null hypothesis that the number of alcoholics in a family has a binomial distribution (like failures in a sequence of independent trials), and a further assumption that every alcoholic has the same independent chance of being admitted to a clinic.

Let π be the probability of an individual becoming an alcoholic, and suppose that the probability that a member of a family becomes an alcoholic is independent of whether another member is alcoholic or not. Further let $p(n)$, $n = 1,2,...$, be the probability distribution of family size (whether a family has an alcoholic or not) in the general population. Then the probability that a family is of size n and has r alcoholics is

$$p(n)\binom{n}{r}\pi^r\phi^{n-r}, \quad r = 0,...,n, \quad n = 1,2,..., \tag{6.2}$$

where $\phi = (1-\pi)$. From (6.2), it follows that the distribution of family size in the general population, given that a family has at least one alcoholic, is

$$\frac{(1-\phi^n)p(n)}{1-E(\phi^n)}, \quad n = 1,2,\ldots . \tag{6.3}$$

If we had chosen households at random and recorded the family sizes in households containing at least one alcoholic, then the null hypothesis on the excess of alcoholics in larger families could be tested by comparing the observed frequencies with the expected frequencies under the model (6.3). However, under the sampling scheme adopted of ascertaining the values of n and r from an alcoholic admitted to a clinic, the weighted distribution of (n,r),

$$p^w(n,r) = rp(n)\binom{n}{r}\frac{\pi^r\phi^{n-r}}{\pi E(n)}, \tag{6.4}$$

is more appropriate. If we had information on the family size n as well as on the number of alcoholics (r) in the family, we could have compared the observed joint frequencies of (n,r) with those expected under the model (6.4).

From (6.4), the marginal distribution of n alone is

$$np(n)/E(n), \quad n = 1,2,\ldots, \tag{6.5}$$

which is used by Smart and Sprott as a model for the observed frequencies of family sizes. It is shown in (6.3) that in the general population, the distribution of family size in families with at least one alcoholic is

$$\frac{(1-\phi^n)p(n)}{1-E(\phi^n)},$$

which reduces to (6.5) if ϕ is close to unity. In other words, if the probability of an individual becoming an alcoholic is small, then the distribution of family size as ascertained is close to the distribution of family size in families with at least one alcoholic in the general population. This is not true if ϕ is not close to unity.

Smart and Sprott found that the distribution (6.5) did not fit the observed

frequencies, which had heavier tails. They concluded that larger families contribute more than their expected share of alcoholics. Is this a valid conclusion? It is seen that the weighted distribution (6.5) is derived under two hypotheses. One is that the distribution of family size in the subset of families having at least one alcoholic in the general population is of the form (6.3) which is implied by the original null hypothesis posed by Smart. The other is that the method of ascertainment is equivalent to p.p.s. sampling of families, with probability proportional to the number of alcoholics in a family. The rejection of (6.5) would imply the rejection of the first of these two hypotheses if the second is assumed to be correct. There are no *a priori* grounds for such an assumption, and in the absence of an objective test for this, some caution is needed in accepting Smart's conclusions.

Another hypothesis considered by Smart was that the later-born children have a greater tendency to become alcoholic than the earlier-born. The method used by Smart may be somewhat confusing to statisticians. Some comments were made by Sprott criticizing Smart's approach. We shall review Smart's analysis in the light of the model (6.4). If we assume that birth order has no relationship to becoming an alcoholic, and the probability of an alcoholic being referred to a clinic is independent of the birth order, then the probability that an observed alcoholic belongs to a family with n children and r alcoholics and has given birth order $s \leq n$ is, using the model (6.4),

$$\frac{rp(n)}{nE(n)}\binom{n}{r}\pi^{r-1}\phi^{n-r}, \quad s = 1,\ldots,n; \quad r = 1,\ldots,n, \quad n = 1,2,\ldots . \tag{6.6}$$

Summing over r, we find that the marginal distribution of (n,s), the family size and birth order, applicable to the observed distribution, is

$$p(n)/E(n), \quad s = 1,\ldots,n, \quad n = 1,2,\ldots, \tag{6.7}$$

where it may be recalled that p(n), n = 1,2,..., is the distribution of family

size in the general population. Smart gave the observed bivariate freqencies of (n,s), and since p(n) was known, the expected values could have been computed and compared with the observed. But, he did something else.

From (6.7), the marginal distribution of birth rank is

$$\frac{1}{E(n)} \sum_{i=r}^{\infty} p(i), \quad r = 1,2,\ldots . \tag{6.8}$$

Smart's (1963) analysis in his Table 2 is an attempt to compare the observed distribution of birth ranks with the expected under the model (6.8) with p(i) itself estimated from data using the model (6.1).

A better method is as follows: from (6.7) it is seen that for given family size, the expected birth order frequencies are equal as computed by Smart (1963) in Table 1, in which case individual Chi-squares comparing the expected and observed frequencies for each family size would provide all the information about the hypothesis under test. Such a procedure would be independent of any knowledge of p(n). But it is not clear whether a hypothesis of the type posed by Smart can be tested on the basis of the available data without further information on the other alcoholics in the family, such as their age, sex, etc.

Table 6 reproduces a portion of Table 1 in Smart (1963) relating to families up to size 4 and birth ranks up to 4. It is seen that for family sizes 2 and 3, the observed frequencies seem to contradict the hypothesis, and for family sizes above 3 (see Smart's Table 1), birth rank does not have any effect. It is interesting to compare the above data with a similar type of data (Table 7) collected by the author on birth rank and family size of the staff members in two departments at the University of Pittsburgh. It appears that there are too many earlier-borns among the staff members, indicating that becoming a professor is an affliction of the earlier born! It is expected that in data of the kind we are considering there will be an excess

of the earlier born without implying an implicit relationship between birth order and a particular attribute, especially when it is age dependent.

Table 6. Distribution of Birth Rank s and Family Size n^a

	$n = 1$		2		3		4	
s	O	E	O	E	O	E	O	E
1	21	21	22	16	17	13.3	11	11.75
2			10	16	14	13.3	10	11.75
3					9	13.3	13	11.75
4							13	11.75

O = observed, E = expected.

Table 7. Distribution of Birth Rank s and Family Size $n \leq 4$ Among Staff Members (University of Pittsburgh)

s	$n = 1$	2	3	4
1	7	14	9	6
2		6	4	2
3			2	0
4				0

7. WAITING TIME PARADOX

Patil (1984) reported a study conducted in 1966 by the Institute National de la Statistique el de l'Economie Appliquee in Morocco to estimate the mean sojourn time of tourists. Two types of surveys were conducted, one by contacting tourists residing in hotels and another by contacting tourists at frontier stations while leaving the country. The mean sojourn time as reported by 3,000 tourists in hotels was 17.8 days, and by 12,321 tourists at frontier stations was 9.0. Suspected by the officials in the department of planning, the estimate from the hotels was discarded.

It is clear that the observations collected from tourists while leaving the country correspond to the true distribution of sojourn time, so that the observed average 9.0 is a valid estimate of the mean sojourn time. It can be shown that in a steady state of flow of tourists, the sojourn time as report-

ed by those contacted at hotels has a size biased distribution, so that the observed average will be an overestimate of the mean sojourn time. If X^w is a size biased random variable (r.v.), then

$$E(X^w)^{-1} = \mu^{-1} \tag{7.1}$$

where μ is the expected value of X, the original variable. The formula (7.1) shows that the harmonic mean of the size biased observations is a valid estimate of μ. Thus the harmonic mean of the observations from the tourists in hotels would have provided an estimate comparable with the arithmetic mean of the observations from the tourists at the frontier stations.

It is interesting to note that the estimate from hotel residents is nearly twice the other, a factor which occurs in the waiting time paradox (see Feller, 1966; Patil and Rao, 1977) associated with the exponential distribution. This suggests, but does not confirm, that sojourn time distribution may be exponential.

Suppose that the tourists at hotels were asked how long they had been staying in the country up to the time of inquiry. In such a case, we may assume that the p.d.f. of the r.v. Y, the time a tourist has been in a country up to the time of inquiry, is the same as that of the product X^wR, where X^w is the size biased version of X, the sojourn time, and R is an independent r.v. with a uniform distribution on [0,1]. If F(x) is the distribution function of X, the the p.d.f. of Y is

$$\mu^{-1}[1 - F(y)]. \tag{7.2}$$

The parameter μ can be estimated on the basis of observations on Y, provided the functional form of F(y), the distribution of the sojourn time, is known.

It is interesting to note that the p.d.f. (7.2) is the same as that obtained by Cox (1962) in studying the distribution of failure times of a component used in different machines from observations of the ages of the

components in use at the time of investigation.

8. DAMAGE MODELS

Let N be a r.v. with probability distribution, p_n, n = 1,2,..., and R be a r.v. such that

$$P(R = r|N = n) = s(r,n). \tag{8.1}$$

Then the marginal distribution of R truncated at zero is

$$p_r' = (1-p)^{-1} \sum_{n=r}^{\infty} p_n s(r,n), \quad r = 1,2,..., \tag{8.2}$$

where

$$p = \sum_{1}^{\infty} p_i s(0,i). \tag{8.3}$$

The observation r represents the number surviving when the original observation n is subject to a destructive process which reduces n to r with probability s'. Such a situation arises when we consider observations on family size counting only the surviving children (R). The problem is to determine the distribution of N, the original family size, knowing the distribution of R and assuming a suitable survival distribution.

Suppose that $N \sim P(\lambda)$, i.e., distributed as Poisson with parameter λ, and let $R \sim B(\cdot,\pi)$, i.e. binomial with parameter π. Then

$$p_r' = e^{-\lambda\pi} \frac{(\lambda\pi)^r}{r!(1 - e^{-\lambda\pi})}, \quad r = 1,2,... . \tag{8.4}$$

It is seen that the parameters λ and π get confounded, so that knowing the distribution of R, we cannot find the distribution of N. Similar confounding occurs when N follows a binomial, negative binomial, or logarithm series distribution. When the survival distribution is binomial, Sprott (1965) gives

a general class of distributions which has this property. What additional information is needed to recover the original distribution? For instance, if we know which of the observations in the sample did not suffer damage, then it is possible to estimate the original distribution as well as the binomial parameter π.

It is interesting to note that observations which do not suffer any damage have the distribution

$$p_r^u = cp_r\pi^r, \tag{8.5}$$

which is a weighted distribution. If the original distribution is Poisson, then

$$p_r^u = e^{-\lambda\pi}\frac{(\lambda\pi)^r}{r!(1-e^{-\lambda\pi})}, \tag{8.6}$$

which is the same as (8.4). It is shown in Rao and Rubin (1964) that the equality $p_r^u = p_r'$ characterizes the Poisson distribution.

The damage models of the type described above were introduced in Rao (1965). For theoretical developments on damage models and characterization of probability distributions arising out of their study, the reader is referred to Alzaid, Rao and Shanbhag (1984).

9. QUADRAT SAMPLING WITH VISIBILITY BIAS

For the purpose of estimating wildlife population density, quadrat sampling has been found generally preferable. Quadrat sampling is carried out by first selecting at random a number of quadrats of fixed size from the region under study and ascertaining the number of animals in each. Following Cook and Martin (1974) we make the assumptions as given below:

A_1: Animals occur in groups within each quadrat and the number of groups within a quadrat has a specified distribution.

A_2: The number of animals in a group has a specified distribution.

A_3: The number of groups within a quadrat and the number of animals within the groups are all independently distributed.

A_4: The method of sampling is such that the probability of sighting (recording) a group of x animals is w(x).

Let X and X^w be the r.v.'s representing the number of animals in a group in the population and as ascertained. Similarly, let N and N^w be the r.v.'s for the number of groups within a quadrat. It is clear that since the method of ascertainment does not give equal chance of selection to groups of all sizes (unless w(x) is constant), the r.v.'s X and X^w do not have the same distribution, and so is the case with N and N^w. The following theorem provides the basic results in quadrat sampling theory.

THEOREM. Under the assumptions A_1-A_4 we have the following results.

(i) $$P(N^w = m | N = n) = \binom{n}{m}\omega^m(1-\omega)^{n-m}$$

where

$$\omega = \sum_1^\infty w(x)P(X = x)$$

is the visibility factor (the probability of recording a group).

(ii) $$P(N^w = m) = \sum_{n=m}^\infty \binom{n}{m}\omega^m(1-\omega)^{n-m}P(N = n),$$

i.e., the visibility bias induces an additive damage model on the true quadrat frequency with binomial survival distribution (see Rao 1965).

(iii) The probability that m observed groups in a quadrat have $x_1, \ldots, x_m$ animals is

$$P(X_1^W = x_1, \ldots, X_m^W = x_m | N^W = m) = \prod_{i=1}^{m} P(X_i^W = x_i)$$

where it may be noted,

$$P(X^W = x) = w(x)P(X = x)/\omega.$$

(iv) Let $S^W = X_1^W + \ldots + X_m^W$. Then

$$P(X^W = y) = \sum_{m=1}^{\infty} P(N^W = m)P(S^W = y|m)$$

$$P(S^W = y|m) = \sum_{\Sigma x_i = y} \frac{w(x_1)}{\omega} \ldots \frac{w(x_m)}{\omega} P(X_1 = x_1) \ldots P(X_m = x_m).$$

Proof. Under the assumptions and notations used, we have the basic probability equation

$$P(N = n, N^W = m, X_1^W = x_1, \ldots, X_m^W = x_m, X_{m+1} = x_{m+1}, \ldots, X_n = x_n)$$

$$= P(N = n)\binom{n}{m} \prod_{j=1}^{m} P(X_j = x_j)w(x_j) \prod_{j=m+1}^{n} [1-w(x_j)]P(X_j = x_j). \qquad (9.1)$$

From (9.1) summing out $X_{m+1}, \ldots, X_n$ we have

$$P(N = n, N^W = m, X_1^W = x_1, \ldots, X_m^W = x_m)$$

$$= P(N = n)\binom{n}{m}\omega^m(1-\omega)^{n-m} \prod_{j=1}^{m} P(X_j^W = x_j). \qquad (9.2)$$

Then the results (i), (ii) and (iii) of the theorem follow from (9.2). Summing (9.2) over n from m to ∞, we have

$$P(N^W = m, X_1^W = x_1, \ldots, X_m^W = x_m) = P(N^W = m) \prod_{j=1}^{m} P(X_j^W = x_j), \qquad (9.3)$$

from which the result (iv) follows.

Note 1. The expression (9.3) enables us to write down the joint likelihood of the numbers of groups observed in different quadrats and the numbers of animals observed in all the groups sighted. Thus, if $m_1, \ldots, m_k$ are the numbers of groups in k quadrats and x_{ij} is the number of animals in the j-th quadrat, the joint likelihod is the product of

$$\prod_{i=1}^{k} P(N^w = m_i) \tag{9.4}$$

and

$$\prod_{i=1}^{k} \prod_{j=1}^{m_i} P(X_{ij}^w = x_{ij}). \tag{9.5}$$

Results (ii) and (iii) of the theorem give the methods of computing the individual terms in (9.4) and (9.5) from the population distributions of N and X and the weight function w(x). In general, the unknown parameters are those occurring in the specified distributions of N and X and the additional visibility factor ω (or p the probability of sighting an animal). All these could be estimated using the product of (9.4) and (9.5) as the likelihood function.

Note 2. Cook and Martin (1974) consider the special case where

$$N \sim P_0(\lambda), \quad \text{Poisson with parameter } \lambda, \tag{9.6}$$

$$X \sim a_x \theta^x / g(\theta), \quad \text{power series distribution}, \tag{9.7}$$

$$w(x) = 1 - (1 - \beta)^x. \tag{9.8}$$

It may be noted that whatever w(x) may be, $N^w \sim P_0(\delta)$, $\delta = \lambda\omega$ where

$$\omega = \Sigma s_x w(x) \theta^x / g(\theta) \quad \text{and} \quad X^w \sim a_x w(x) \theta^x / \omega g(\theta).$$

Thus, there are three parameters δ, ω and θ. Then the parameter δ is estimated

from the likelihood (9.4) and ω, θ from (9.5). Cook and Martin (1974) provided the necessary computations in such a case, choosing $w(x)$ as in (9.8).

If N is not a Poisson variable, then the distribution of N^w involves ω as an additional parameter (see Rao (1965) and Sprott (1965)), in which case the product of (9.4) and (9.5) provide the joint likelihood for the estimation of all the unknown parameters.

In the special case where N and X are as distributed in (9.6) and (9.7) respectively and $w(x) = \beta^x$ (i.e., when a group is observed if and only if all the animals are sighted),

$$N^w \sim P_0(\delta),\ \delta = \lambda\omega \quad \text{and} \quad X^w \sim a_x\phi^x/g(\phi),\ \phi = \theta\beta$$

so that the parameters λ, θ and β are confounded and are not individually estimable.

10. THE STORY OF BROKEN BONES

The following problem arose in the analysis of measurements on femur bones recovered from an ancient graveyard. When a femur bone was found intact it was possible to take three measurements, length L, breadth of the top tip B and breadth of the bottom tip T. But when a broken piece was found, either the measurement B or the measurement T could be taken. Thus, the observed data was incomplete with either the measurement B alone or T alone on some and all three L, B, T on others. How does one estimate from the fragmentary data of the above type the mean values and second order moments of L, B, T in the original population of femur bones?

Let $p(\ell,b,t)$ be the p.d.f. of L, B, T in the original population with the associated marginal densities

$$p(b) = \int p(\ell,b,t)d\ell dt \quad \text{and} \quad p(t) = \int p(\ell,b,t)d\ell db. \tag{10.1}$$

If the probability that a bone gets broken does not depend on its dimensions, then the likelihood of the observed data could be written down using the p.d.f.'s, $p(\ell,b,t)$, $p(b)$ and $p(t)$, depending on the available measurements on each specimen. However, it may happen that the longer bones have a greater chance of being broken; such a phenomenon was demonstrated in a similar situation on skull measurements by Rao and Shaw (1948) and Rao (1978). In such a case we may have to distinguish the measurements L^s, B^s, T^s taken on well preserved (surviving) bones and measurements L^d, B^d, T^d associated with the damaged bones and denote their p.d.f.'s with superfixes s and d respectively.

We suppose that the chance of survival of a femur bone of length ℓ is $s(\ell)$ depending only on ℓ. Then

$$p^s(\ell,b,t) = \sigma^{-1}p(\ell,b,t)s(\ell), \quad \sigma = E[s(\ell)]. \tag{10.2}$$

Similarly

$$p^d(\ell,b,t) = (1-\sigma)^{-1}p(\ell,b,t)\big(1-s(\ell)\big). \tag{10.3}$$

From (10.2) and (10.3), the following are immediately deduced:

$$p^s(\ell) = \sigma^{-1}p(\ell)s(\ell),$$

$$p^s(b,t|\ell) = p(b,t|\ell),$$

$$p^s(b,t) = \int \sigma^{-1}p(b,t|\ell)p(\ell)s(\ell)d\ell$$

$$= \int \sigma^{-1}p(b,t)p(\ell|b,t)s(\ell)d\ell = p(b,t)w(b,t),$$

$$p^s(\ell|b,t) = \frac{p(\ell,b,t)s(\ell)}{\int p(\ell,b,t)s(\ell)d\ell} \neq p(\ell|b,t),$$

$$p^d(b,t|\ell) = p(b,t|\ell),$$

$$p^{s \text{ or } d}(b) = p(b) \quad \text{and} \quad p^{s \text{ or } d}(t) = p(t),$$

$$p^{d}(\ell|b,t) = \frac{p(\ell,b,t)\big(1-s(\ell)\big)}{\int p(\ell,b,t)\big(1-s(\ell)\big)d\ell} \neq p(\ell|b,t).$$

It is interesting to note that all distributions involving L as a main variate are weighted. One casualty of this result is that the regression of L on (B,T) estimated from the complete sets of samples on L, B, T does not correspond to the true regression of L on (B,T) in the original population of femur bones. But others like

$$p^{s}(b,t|\ell), \quad p^{d}(b,t|\ell), \quad p^{s \text{ or } d}(b), \quad p^{s \text{ or } d}(t) \tag{10.4}$$

are independent of $s(\ell)$, and the properties of these distributions could be used to estimate all the unknown parameters when $s(\ell)$ is unknown.

For instance using all the available measurements on B and T (both on damaged and well preserved bones), the mean values μ_B and μ_T of B and T in the original population could be estimated by the usual averages. From the observations on the complete set of L, B and T we can estimate the regressions of B on L and T on L in the usual way. Then the missing values of L can be estimated in each case, i.e., where B alone or T alone is available, by inverse regression using the regression equation of B on L or T on L. Now the average of the observed values of L and the estimated values of L in missing cases is computed as an estimate of μ_L, the mean value of L in the original population. In a similar manner the second order moments can be estimated using the relationships between the parameters of the original distribution of L, B and T and of the conditional distributions (10.4).

13. CALCUTTA BLACKOUT DISTRIBUTION

Suppose that we are conducting an experiment to measure the time taken for a certain event to happen, and for running the experiment a continuous supply of electric power is needed. If the power supply is cut off before the event happens, then the experiment has to be abandoned and no observation gets recorded. What distribution do the recorded observations resulting only from the successful experiments (when the power supply is on until the event occurred) obey?

Let $f(x)$ be the p.d.f. of X, the time taken for an event to happen, and $g(t)$ be the p.d.f. of T, the time at which the electric supply may fail (in Calcutta this is a random phenomenon producing a blackout). An observation on X gets recorded only when a pair (x,t) occurs such that $x \leq t$. The p.d.f. of a pair (X,T) such that $X \leq T$ is

$$\frac{f(x)g(t)}{P(X \leq T)} \tag{11.1}$$

so that the p.d.f. of the recorded variable $X^{(r)}$ is

$$f^{(r)}(x) = \int_x^\infty \frac{f(x)g(t)}{P(X \leq T)}\,dt = \frac{f(x)\big(1-G(x)\big)}{P(X \leq T)} \tag{11.2}$$

where $G(t)$ is the distribution function of T. The distribution (11.2) is a weighted version of the distribution of X, which I termed as the Calcutta Blackout Distribution (CBD).

If we have observations from successful experiments alone, then the relevant distribution is (11.2). However, in such a situation other observations could be made. The appropriate distributions when additional information is available are discussed below.

If we define a random variable $Z = \min(X,T)$, then it is observable in each experiment. The p.d.f. of Z is

$$h(z) = -\frac{d}{dz}\int_z^\infty \int_z^\infty f(x)g(t)dxdt = [1 - F(z)]g(z) + [1 - G(z)]f(z) \qquad (11.3)$$

which is a mixture of weighted distributions.

In the experiment described above, there is also the possibility of recording $Z_* = \min(X,T)$ with the identifying symbol whether the true observation is on X or T. In such a case the p.d.f. of Z_* is

$$h_*(z) = \begin{cases} f(z)\big(1-G(z)\big) & \text{if} \quad z \text{ is an observation on } X, \\ g(z)\big(1-F(z)\big) & \text{if} \quad z \text{ is an observation on } T. \end{cases} \qquad (11.6)$$

12. CLOUDED DISTRIBUTIONS

When the sea-surface temperature is measured by a satellite, there is a possibility that the reading is effected by a cloud cover resulting in reduced values of temperatures. The amount by which a measurement is scaled down depends on the thickness of the cloud. But when a large number of measurements are taken in a given area, there will be a proportion of data which is free from cloud contamination while the rest are effected by clouds of different thickness. If p(x) is the true distribution of the sea-surface temperatures, whose average we are seeking, q(c), $0 \le c \le 1$, is the p.d.f. of cloudiness in the area under cloud cover and λ is the proportion of the area without cloud cover, then the p.d.f. relevant to the observed temperatures is

$$\lambda p(t) + (1-\lambda)\int \frac{1}{c} p(\frac{t}{c})q(c)dc. \qquad (12.1)$$

The proportion λ and the p.d.f. q(c) are generally unknown in any given situation, and modelling the entire data for the unknown elements is extremely difficult. However, when λ is large, the distribution (12.1) is dominated by p(t) in the right tail, and this can be judged by the smoothness of the

histogram of the observed data relating to large values of the temperature. When this happens, we can consider the data in the tails of the histogram as uncontaminated observations and use them only in estimating the mean sea-surface temperature. Such a technique was used by Smith, Rao, Koffler and Curtis (1970). They assumed that the temperature distribution is normal (with mean μ and variance σ^2) and an estimate of σ^2 is available from a different source, and equated the observed (estimate) point of inflexion in the right tail of the smoothed histogram to $\mu + \sigma$, which provided an estimate of μ. An alternative method is to consider a truncation point τ and estimate the mean using only the observations which are equal to or exceeding τ. The estimate of μ in such a case satisfies the equation

$$\bar{t}_\tau = \mu + \frac{\sigma Z(\frac{\tau-\mu}{a})}{1 - \Phi(\frac{\tau-\mu}{a})} \qquad (12.2)$$

where $\bar{t}_\tau$ is the average of observations greater than or equal to τ. We denote the solution of (12.2) by $\hat{\mu}_\tau$. Then we draw the graph of $\hat{\mu}_\tau$ against τ and choose that value of τ, say τ_0, from where the graph shows a tendency to be parallel to the τ axis. The estimate of μ is taken as $\hat{\mu}_{\tau_0}$.

13. REFERENCES

[1] ALZAID, A.H., RAO, C.R. and SHANBHAG, D.N. (1984). Solutions of certain functional equations and related results on probability distributions. Technical Report, University of Sheffield, U.K.

[2] COOK, R.D. and MARTIN, F.B. (1974). A model for quadrat sampling with visibility bias. *J. Amer. Statist. Assoc.*, 69, 345-349.

[3] COX, D.R. (1962). *Renewal Theory.* Chapman and Hall, London.

[4] FELLER, W. (1966). *An Introduction to Probability Theory and its Applications,* Vol. 2. John Wiley & Sons, New York.

[5] FELLER, W. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1 (3rd edn.). John Wiley & Sons, New York.

[6] FISHER, R.A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.*, 6, 13-25.

[7] PATIL, G.P. 1984). Studies in statistical ecology involving weighted distributions. In *Statistics: Applications and New Directions*, pp.478-503. Indian Statistical Institute, Calcutta.

[8] PATIL, G.P. and ORD, J.K. (1976). On size-biased sampling and related form-invariant weighted distributions. *Sankhyā Ser. B*, 38, 48-61.

[9] PATIL, G.P. and RAO, C.R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, (P.R. Krishnaiah, Ed.), pp.383-405. North-Holland Publishing Company, Amsterdam.

[10] PATIL, G.P. and RAO, C.R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179-189.

[11] RAO, C.R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions*, (G.P. Patil, Ed.), pp.320-333. Statist. Publishing Society, Calcutta. Reprinted in *Sankhyā Ser. A*, 27, 311-324.

[12] RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, (2nd edn.). John Wiley & Sons, New York.

[13] RAO, C.R. (1975). Some problems of sample surveys. *Suppl. Adv. Appl. Probab.*, 7, 50-61.

[14] RAO, C.R. (1977). A natural example of a weighted binomial distribution. *Amer. Statist.*, 31, 24-26.

[15] RAO, C.R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In *A Celebration of Statistics*, the ISI Centenary Volume, (A.C. Atkinson and S.E. Fienberg, Eds.), pp.543-569. Springer-Verlag.

[16] RAO, C.R. and RUBIN, H. (1964). On a characterization of the Poisson distribution. *Sankhya Ser. A*, 25, 295-298.

[17] RAO, C.R. and SHAW, D.C. (1948). On a formula for the prediction of cranial capacity. *Biometrics*, 4, 247-253.

[18] SMART, R.G. (1963). Alcoholism, birth order, and family size. *J. Abnorm. Soc. Psychol.*, 66, 17-23.

[19] SMART, R.G. (1964). A response to Sprott's "Use of Chi-square". *J. Abnorm. Soc. Psychol.*, 69, 103-105.

[20] SMITH, W.L., RAO, P.K., KOEFFLER, R. and CURTIS, W.P. (1970). The determination of sea-surface temperature from satellite high resolution infrared window radiation measurements. *Monthly Weather Review*, 98, 604-611.

[21] SPROTT, D.A. (1964). Use of Chi-square. *J. Abnorm. Soc. Psychol.*, 69, 101-103.

[22] SPROTT, D.A. (1965). Some comments on the question of identifiability of parameters raised by Rao. In *Classical and Contagious Discrete Disturbances*, (G.P. Patil, Ed.), pp.333-336. Statist. Publishing Society, Calcutta.